Discerning Emotions of Bloggers based on Topics – a Supervised Coreference Approach in Bengali

Dipankar Das
Department of Computer Science and Engineering
Jadavpur University
dipankar.dipnil2005@gmail.com

Sivaji Bandyopadhyay
Department of Computer Science and Engineering
Jadavpur University
sivaji_cse_ju@yahoo.com

Abstract

This paper presents an approach to identify the users' emotions on different topics provided in Bengali blog posts. The identification of emotion holder, emotion topic along with emotional expression aims to develop the baseline system. The input vectors containing emotional expressions and topics with respect to the corresponding emotion holders are prepared from the annotated blog posts. The feature vectors consist of different syntactic, semantic, rhetoric and overlapping features are employed in a supervised system to identify the co reference of the emotion holder with the corresponding emotional expression and topic. Evaluation metric (*Krippendorff's* α) achieves the agreement scores of 0.53 and 0.67 for the baseline and supervised coreference classification systems respectively. The emotions for different topics with respect to each of the blog users represent the topic dependent users' emotional views.

Keywords: Emotional Expression, Holder, Topic, coreference, Emotional View.

[1. Introduction]

Major studies on Opinion Mining and Sentiment Analyses have been attempted with more focused perspectives rather than fine-grained emotions. In psychology and common use, emotion is an aspect of a person's mental state of being, normally based in or tied to the person's internal (physical) and external (social) sensory feeling [21]. The source or holder of an emotional expression is the speaker or writer or experiencer [17]. Extraction of emotion holder is important for discriminating between emotions that are viewed from different perspectives [22]. By grouping opinion holders of different stance on diverse social and political issues, we can a have better understanding of the relationships among countries or among organizations [11]. Topic is the real world object, event, or abstract entity that is the primary subject of the emotion or opinion as intended by the holder and the topic depends on the context in which its associated emotional expression occurs [15]. The following Bengali sentence shows the emotional expression, its associated holder and topic. As the sentence is collected from a blog post, the *writer* is also considered as a default holder [17].

Example 1:

*Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), Pages 350-363,
Puli, Nantou,Taiwan, September 2010.*

রাশেদ বলেছেন আপনার **কবিতাটা** পড়তে গিয়ে
(*Rashed*) (*bolechen*) (*apnar*) (*kobitata*) (*porte*) (*giye*)
তার এই **সুন্দর** **কৌতুকটা** মনে পড়ছিলো।
(*tar*) (*ei*) (*sundar*) (*koutukta*) (*mone*) (*porchilo*).
Rashed *said that he was remembering this* ***beautiful comic*** *while reading your* ***poem****.*

Emotional Expression: সুন্দর কৌতুক (***beautiful comic***), *Holder:* < writer**,** রাশেদ (***Rashed***) **>,** *Topic:* কবিতা (***poem***)

The information of emotion is useful for the domain of Question Answering (QA), Information Retrieval (IR), product reviews, social media, stock markets, and customer relationship management. Especially, the blog posts contain instant views, updated views or influenced views regarding single or multiple topics. Blogs are the communicative and informative repository of text based emotional contents in the Web 2.0 [2]. Researches on emotion show that blogs play the role of a substrate to analyze the reactions of different emotional enzymes. Many blogs act as online diary of the blogger reporting on the blogger's daily activities and surroundings. Sometimes, blog posts are annotated by other bloggers. Large blog data set is also suitable for machine learning models.

Thus the present task deals with the identification of users' emotion on different topics from a Bengali blog corpus [7]. Each sentence of the Bengali blog corpus is annotated with the emotional components such as emotional expression (word/phrase), intensity, associated holder and topic(s). Ekman's six emotion classes (*anger, disgust, fear, happy*, *sad* and *surprise*) along with three types of intensities (*high*, *general* and *low*) are annotated at sentence level. The 3367 sentences with respect to eight different potential topics along with 22 different blog users are considered for conducting the present task. The emotion topic annotated in each of the sentences is directly linked with the topic of the document that contains the sentence.

With the above examples and problems in mind, we hypothesize that the notion of user-topic coreference will facilitate both the manual and automatic identification of emotional views: Two components of emotion such as holders and topics are emotion coreferent if they share the same emotional expressions. The baseline system extracts the emotional expressions using Bengali WordNet Affect lists [6]. An unsupervised system is developed for identifying emotion holders and topics with respect to the emotional expressions. The co reference of the emotion topics and emotion holders is measured using Passonneau's (2004) [24] generalization of Krippendorff's (1980) [25] α metric.

On the other hand, The Support Vector Machine [10] based supervised classifier is employed for coreference classification. Each of the input vectors containing emotional expression, associated holder and topic is prepared from each of the annotated Bengali blog sentences. The feature vector is prepared based on the information present in the sentences containing lexical, syntactic, semantic, rhetoric and overlapping features (word, part-of speech (POS), Named Entity (NE)). Training with 2234 sentences, the feature analysis has been conducted on the development set of 630 sentences. The standard Krippendorff's (1980) [25] α metric produces a score of 0.67 that significantly outperforms the baseline with a score of 0.53 on the test set of 503 sentences. The observation suggests that, the rhetorical structure improves the performance of the coreference classification reasonably. The classification produces errors in resolving the overlapping context that contains the emotional expression and topic.

The rest of the paper is organized as follows. Section 2 describes the related work. The baseline system is described in Section 3. The supervised framework with features is discussed in Section 4. Evaluation results along with feature analysis and error reducing mechanisms are specified in Section 5. Finally Section 6 concludes the paper.

[2. Related Work]

In order to estimate affects in text, the model proposed in [19] processes symbolic cues and employs natural language processing techniques for word/phrase/sentence level analysis, considering relations among words in a sentence. The current trend in the emotion analysis area is exploring machine learning techniques [28], which consider the problem as text categorization or analogous to topic classification that underscores the difference between machine learning methods and human-produced baseline models [29]. Affective text shared task on news headlines at SemEval 2007 for emotion and valence level identification [20] has drawn the focus to this field.

Prior work in identification of opinion holders has sometimes identified only a single opinion per sentence [30], and sometimes several [1]. Identification of opinion holders for Question Answering with supporting annotation task was attempted in the very beginning [17]. Before that, another work on labeling the arguments of the verbs with their semantic roles using a novel frame matching technique was carried out in [31]. Based on the traditional perspectives, another work discussed in [35] uses an emotion knowledge base for extracting *emotion holder*. The machine learning based classification task for “not holder”, “weak holder”, “medium holder”, or “strong holder” is described in [34]. Kim and Hovy [11] identified opinion holder with topic from media text using semantic role labeling. An anaphor resolution based opinion holder identification method exploiting lexical and syntactic information from online news documents was attempted in [36]. The syntactic models of identifying *emotion holder* for English emotional verbs are developed in [5].

In the related area of opinion topic extraction, different researchers contributed their efforts. Some of the works are mentioned in [13] [18] [37]. But, all these works are based on lexicon look up and are applied on the domain of product reviews. The topic annotation task on the MPQA corpus is described in [15]. The authors have pointed out that the target spans alone are insufficient for many applications as they neither contain information indicating which opinions are about the same topic, nor provide a concise textual representation of the topics.

The method of identifying an opinion with its holder and topic from online news is described in [11]. The model extracts opinion topics for subjective expressions signaled by verbs and adjectives. They have extracted the topics associated with a specific argument position based on verb or adjective. Similarly, the verb based argument extraction and associated topic identification is considered in the present system.

Opinion topic identification differs from topic segmentation [1]. The opinion topics are not necessarily spatially coherent as there may be two opinions in the same sentence on different topics, as well as opinions that are on the same topic separated by opinions that do not share that topic [15]. The hypothesis is established by applying the technique of coreference classification for topic annotation. The building of fine-grained topic knowledge based on rhetorical structure and segmentation of topics using different types of lexical, syntactic and overlapping features substantially reduces the problem of emotion topic distinction in our present supervised framework.

Moreover, all the above-cited works have been attempted for English. Recent study shows that non-native English speakers support the growing use of the Internet[1]. In addition to that, rapidly growing web users from multilingual communities focus the attention to improve the multilingual search engines on the basis of sentiment or emotion This raises the demand of emotion analysis for languages other than English. Bengali is the sixth popular language in the World[2], second in India and the national language in Bangladesh but it is less computerized compared to English. Works on emotion analysis in Bengali have started recently [2] [3]. The comparative evaluation of the features on equivalent domain for Bengali and English language can be found in [4]. To the best of our knowledge, at present, there is no such user-topic coreference analysis of emotion has been conducted for Bengali or even for other Indian languages. Thus we believe that this work would meet the demands of topic focused emotion analysis systems.

[3. Baseline System]

Identifying Emotion Expression: The sentences are passed through an open source Bengali shallow parser[3]. The shallow parser gives different morphological information (root, lexical category of the root, gender, number, person, case, vibhakti, tam, suffixes etc.) that help in identifying the lexical patterns of emotional expressions. The shallow parsed sentences are pre-processed to generate the simplified patterns. The lexical pattern of the shallow parsed result of Example 1 is shown in Figure 1. We extract all component words from the chunks that contain at least one emotion word (e.g. কৌতুক *koutuk* ‘comic’), and match them against the Bengali *WordNet Affect* lists [6]. The words present in the extracted chunks are then treated as candidate seeds for the anchoring vectors representing emotional expressions. Identification of an emotional expression containing a single emotion word is straightforward. But, we include the all the words that present in the chunks in order to identify long emotional expressions. Consecutive words that appear in the chunks and contain at least one emotion word are identified as an emotional expression.

```
((JJP      সুন্দর                         JJ      <fs      af ='সুন্দর
,adj  ,,,,d ,শুন্য,শুন্য'>    )
( NP কৌতুকটা   NN           <fs af =' কৌতুক ,v ,,,,, টা , টা > ))
```

[Figure 1. Example of a pre-processed shallow parsed result]

In many cases, the components of a given emotional expression are separated by stop words (e.g. একটি *ekti* ‘a’, ঐ *oi* ‘that’, এই *ei* ‘this’) conjuncts (e.g. এবং *ebong* ‘and’, অথবা *athoba* ‘or’, কিন্তু *kintu* ‘but’ etc.), negations (e.g. নয় *noy* ‘not’, না *na* ‘neither’ etc.) or intensifiers (তাই *tai* ‘so’, খুব *khub* ‘very’, কম *kam* ‘less’, বেশি *beshi* ‘much’). The aim is to accumulate the words for constructing the emotional expressions with maximum coverage of the contributory components. Each emotional expression is tagged with the Bengali *WordNet Affect* classes in which the words of the emotional expression occur. Each of the sentences is also tagged with Ekman’s six emotion tags [8] that are associated with its corresponding emotional expressions. Each sentence may contain more than one emotion tag. Otherwise, the sentences are treated as neutral sentences.

[1] http://www.internetworldstats.com/stats.htm

[2] http://www.ethnologue.com/ethno_docs/distribution.asp?by=size

[3] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

Identifying Emotion Holder: The baseline model considers the phrasal pattern containing similarity clues to identify emotion holders. The patterns are grouped according to the part-of-speech (POS) categories. It is observed that the hints are present mostly in the user comment portions of the Bengali blogs. Each of the user comment portions started with a corresponding username. The username is the default hint that helps in capturing the first holder present in an anchoring vector of nested sources.

The named entities that are tagged with NNPC (Compound proper noun), NNP (Proper noun), NNC (Compound common noun), NN (Common noun) or PRP (Pronoun) present at the beginning of a sentence are tagged as the responsible candidates of *emotion holder*. The similarity pattern consists of two phrasal constituents, the subject and the verb. The common portions (*Common_Portion*) containing the additional constituents are basically the floating portions. As Bengali is a free phrase order language, the ordering between the verb and the floating portion is not fixed.

The general POS level pattern such as [<NNP/NNPC/NN/NNC/PRP> {<VBZ/VM><*Common_Portion*>}] is considered for capturing the clue of an *emotion holder*. The components of the common portion are assembled starting with the hint of the first occurring POS tags of types NNP or NNC or PRP in the POS tagged sentence. Reaching at the verb based POS tags like VBZ or VM, the system stops incorporating the components into the common portion. The rest of any component present in the sentence after the verb is therefore simply appended to finally build the common portion.

The similarity patterns exist mostly in the simple sentences. The complex or compound sentences are hard to classify into this category. As the identification of *emotion holder* from long complex or compound sentences shows problems in the baseline system, the system fails to identify the nested *emotion holder*s.

Identifying Emotion Topic: The shallow chunked portions formed by removing emotional expressions and holders are identified as the responsible spans that contain one or more potential emotion topics. The words of the shallow chunks containing POS tags of NNP or NNC are allowed to include as emotion topics.

User-Topic Coreference: The identified holders and topics associated with an emotional expression in each of the sentences are coreferent if they share the same emotional expression. The emotion topic is intended by the emotion holder and the topic depends on the context in which its associated emotional expression occurs [16]. Based on the hypothesis, a rule based unsupervised technique is devised to identify the coreference between user and topic with respect to a particular type of emotional expression if the chunks responsible for emotion holder or topic are the immediate neighboring chunks of the emotional expression.

[4. Supervised Framework]

Topic coreference resolution resembles another well-known problem in NLP - noun phrase (NP) coreference resolution that considers machine learning frameworks [26] [27]. Therefore, we adapt a standard machine learning based approach to user-topic coreference resolution.

The Support Vector Machine (SVM) [10] based supervised framework has been used to extract the input vectors that contain emotional expressions, users or holders and topics. The system is trained with 2234 sentences. The best feature set is identified using the 630 development sentences. The Information Gain Based Pruning (IGBP), Admissible Tag Sequence (ATS), Class Splitting technique and Emotional Composition features are applied on the development set and it improves the performance of the supervised system significantly. The detail results are shown in Table 1. All the results are obtained by 10 fold cross validation method.

Feature plays a crucial rule in the SVM framework. By manually reviewing the blog data and different language specific characteristics, word level as well as context level features have been selected heuristically for our classification task. The heart of our method is a pairwise user-topic coreference classifier. Given a pair of emotion holder and topic (and their associated emotional expression), the goal of the classifier is to determine whether the holder and topic are emotion-coreferent. We use the manually annotated data to automatically learn the pairwise classifier. We construct each training example for every pair of emotion holder and topic in the document (each pair is represented as a input vector). The pair is labeled as a positive example if the holder and topic belong to the same emotional expression and a negative example otherwise. Pair wise coreference classification relies critically on the expressiveness of the features used to describe the user-topic pair. We use the following four categories of features: lexical, syntactic, semantic, rhetoric and overlapping features.

Lexical Features

Parts-of-Speech **(POS)**: We are interested with the *noun*, *adjective*, *verb* and *adverb* words as these are emotion informative constituents. The POS features are extracted from the shallow parsed results.

Negations **(NEG)**: Negative words that are annotated in the corpus (নয় *noy* ‘not’, না *na* ‘neither’ etc).

Conjuncts **(CONJ)**: The annotated *Conjuncts* features (e.g. এবং *ebong* ‘and’, অথবা *athoba* ‘or’, কিন্তু *kintu* ‘but’ etc).

Punctuation Symbols **(Sym)**: Symbols such as comma (,), (!), (?) are often used in single or multiple numbers to emphasize emotional expressions and considered as crucial clues for identifying emotional presence in a sentence.

Emoticons **(emot_icon)**: The emoticons (☺,☹,😐) and their consecutive occurrence generally contribute as much as real sentiment to the emotional expressions that precede or follow it.

Syntactic Features

We augment the knowledge of subcategoriztion frames or syntactic frames in case of identifying emotion holders and topics but the identification of syntactic frames is not straight forward.

Verb Identification: To identify the simple verbs from the shallow parsed or chunked corpus, the words that are tagged as main verb (VM) and belong to the verb group chunk (VGNF) in the corpus are identified (e.g. ভালোবাসা *bhalobasa* ‘love’). For compound or conjunct verbs, the pattern such as {[XXX] (NN) [YYY] (VM)} are retrieved from the shallow parsed corpus (e.g. VGNF {[আনন্দ *ananda*] (NN) [করা *kara*] (VM)} means *enjoy*). The light verb [YYY] tagged with ‘VM’ generally occurs in any inflected form. Different suffixes may be attached to a simple verb or light verb depending on various features like Tense, Aspect, and Person. A Bengali stemmer with an accuracy of 97.09% uses a suffix list to identify the stem form of the Bengali simple verbs and light verbs. Another knowledge base stores the stem forms and the corresponding dictionary forms of 374 Bengali verbs containing simple and light verb entries. The dictionary forms of the Bengali compound or conjunct verbs are made by incorporating the dictionary forms of the light verbs with their preceding noun words that are tagged as 'NN' present in the retrieved lexical patterns.

English Equivalent Synset Identification: The determination of equivalent English verbs of a Bengali verb is carried out using a Bengali to English bilingual dictionary[4]. The method to extract the English equivalent synsets of the Bengali verbs is based on the work done in [32]. We have identified the equivalent English verb synsets of the Bengali verb entries that are present in the dictionary. For example, the dictionary entries for simple verb ভালোবাসা *bhalobasa* 'love' and conjunct verb আনন্দ *ananda* করা *kara* 'enjoy' are as follows.

< ভালোবাসা [bhālōbāsā] **v** to ***love***, to be amorous to wards; to ***like***; to have attachment or affection, fondness for …>

< আনন্দ করা **v**. to ***rejoice***; to ***make merry***….>

Different synonyms for a Bengali verb having the same sense are separated using "," and different senses are separated using ";" in the dictionary. The synonyms including similar senses of the target verb are extracted from the dictionary and yield a resulting set called English Equivalent Synset (EES). For example, two English Equivalent Synsets (EES) are extracted for the conjunct verb আনন্দ *ananda* করা *kara* 'enjoy'.

English Equivalent Frame Identification: It is found that each of the English Equivalent Synsets (EES) occurs in each separate class of English VerbNet [12]. VerbNet associates the semantics of a verb with its syntactic frames and combines traditional lexical semantic information such as *thematic roles* and *semantic predicates* with *selectional restrictions*. Member verbs in the same VerbNet class share common syntactic frames, and thus they are believed to have the same syntactic behavior. The VerbNet files containing member verbs and possible subcategorization frames are stored in XML file format. Hence, the XML files of VerbNet are pre-processed to build up a general list that contains all verbs, their classes and possible subcategorization frames (primary as well as secondary). This pre-processed list is searched to extract the present subcategorization frames for each verb (e.g. *love*) of the English Equivalent Synsets (EES) (e.g. love) corresponding to the Bengali verb. These extracted subcategorization frames are believed to be the valid set of argument structures for the Bengali verbs [32] [33].

Frame Matching: On the other hand, the chunked Bengali sentences are passed through a rule based *phrasal-head* extraction module to identify the phrase level argument structures of the sentences corresponding to the position of the verbs. The extracted *head part* of every phrase from the chunked sentence is considered as a component of the argument structure of that sentence. If the acquired argument structure for a Bengali emotional sentence is matched with any of the available extracted frames of English VerbNet, the *thematic role* based holder information (e.g *Experiencer, Agent, Actor, Beneficiary* etc.) and topic information (e.g *Topic*, *Theme*, *Event* etc.) associated with the English frame syntax is mapped to the appropriate slot of the acquired Bengali argument structure. Tag conversion routines are developed to transform the POS of the system-generated argument structures into the POS of the VerbNet frames.

For example, in simple sentences, the occurrence of the NNPC, NNP, NNC or NN tags preceded by the PRP (Pronoun) NNP, NNC, NN or NNPC tags (may contain case markers (e.g. কে [-*ke*]) and followed by a verb gives similar frame syntax as of the "Basic Transitive"

[4] http://home.uchicago.edu/~cbs2/banglainstruction.html

frame of English VerbNet.

Example 2: Vector: < EH_ রাম , ET_ সিতা >

রাম সিতাকে ভালবাসে

(***Ram***) (*Sitake*) (*bhalobase*)

Ram loves Sita.

Acquired Argument Structure: [NNP NNP-*ke* VM]

Extracted VerbNet Frame Syntax: [<NP value="*Experiencer*" ></VERB><NP-*theme*>]

Example 3.*Vector*: < EH_রাশেদ, EH_*রাম*, ET_সুখ >

রাশেদ অনুভব করেছিল যে রামের সুখ অন্তহীন

(*Rashed*) (*anubhob*) (*korechilo*) (*je*) (***Ramer***) (*sukh*) (*antohin*)

Rashed felt that Ram's pleasure is endless.

The acquired argument structure of another Bengali emotional sentence shown in Example 3 is as follows.

Acquired Argument Structure: [NNP VM DET-*je* S]

The argument structure contains sentential complement "S" started by যে –*je* with DET type POS. The argument structure is acquired for the Bengali verb অনুভব *anubhab* করা *kara* 'feel'. One of the extracted VerbNet frame syntax containing –*that* type sentential complement for the equivalent English verb *feel* is as follows.

Extracted VerbNet Frame Syntax: [<NP value="*Experiencer*" ></VERB>< S-*that* (Sentential –*that* Complement)>]

As the acquired argument structure matches with the extracted VerbNet frame syntax, the *emotion holder* related roles (e.g. *Experiencer*) associated with the VerbNet frame is mapped to the equivalent phrase of the acquired argument structure of the Bengali sentence. The phrase is now considered as the candidate of *emotion holder*. Additionally, the sentence present in the portion of sentential complement is passed by the syntactic model to obtain any implicit *emotion holder* if presents. The appositive cases are identified by considering the suffixes determined by the Bengali stemmer.

The case markers in Bengali are required to identify the *emotion holder*s as the case markers give the useful hints to capture the *selectional restrictions* that play a key role in distinguishing the *emotion holder*s from other valid alternatives.

The XML files of VerbNet are pre-processed to build up a general list that contains all member verbs and their available syntactic frames with holder and topic related *thematic* information (e.g. *Experiencer, Agent, Actor, Beneficiary* and *Topic*, *Theme*, *Event* etc.). The pre-processed list is searched to acquire the syntactic frames of each verb.

Semantic Features

Emotion/Affect Words **(EW)**: The presence of a word in the *WordNet Affect* lists [6] identifies the emotion/affect words.

Intensifiers **(INTF)**: The Bengali *SentiWordNet* is being developed by replacing each word entry in the synonymous set of the English *SentiWordNet* [9] by its possible set of Bengali synsets using a synset based English to Bengali bilingual dictionary being developed as part of the EILMT project[5].

The chunks containing JJ (adjective) and RB (adverb) tagged elements are considered as intensifiers. If the intensifier is found in the *SentiWordNet*, then the positive and negative scores of the intensifier are retrieved from the *SentiWordNet*. The intensifier is classified into the list of positive (pos) (**INTF*pos***) or negative (neg) (**INTF*neg***) for which the average retrieved score is higher.

Multiword Expressions: *Reduplication* (সন্দ সন্দ *sanda sanda* [doubt with fear]) and *Idioms* (তাসের ঘর *taser ghar* [weakly built, গৃহদাহ *grrihadaho* [family disturbance]) are considered as features.

Rhetoric Features

Instead of identifying rhetorical relations [14], the present task acquires the rhetorical components such as *locus*, *nucleus* and *satellite* from a sentence as these rhetoric clues help in identifying the individual topic spans associated in a target span of the sentences. The topic of an opinion depends on the context in which its associated opinion expression occurs [16]. The part of the text span containing annotated emotional expression is considered as *locus*. Primarily, the separation of *nucleus* from *satellite* is done based on the punctuation markers (,) (!) (?). Frequently used *discourse markers* (যেহেতু *jehetu* 'as', যেমন *jemon* 'e.g.', কারণ *karon* 'because', মানে *mane* 'means') and *causal verbs* (ঘটায় *ghotay* 'caused') are also the useful clues if they are explicitly specified in the text.

If any word in the annotated emotional expression co-occurs with any word element of the *nucleus* or *satellite* in the same chunk, the feature is considered as *common rhetoric similarity.* Otherwise, the feature is considered as *distinctive rhetoric similarity*. This features aims to separate emotion topics from non-emotion topics as well as to separate the overlapping possibilities of discrete emotion topic spans from non-topical contiguous regions.

Overlapping Features

Word Overlap: True if the two topic spans contain any contain words in common.

Part-of-Speech Overlap: The *verb, noun*, *adjective* and *adverb* are the informative constituents.

NP Coreference: True if the two chunk spans contain NPs that are determined to be coreferent by a simple rule based coreference system.

Named Entity (NE): Each of the sentences is passed through a Named Entity Recognizer [38] for identifying the named entities. If any word is tagged as a named entity and present in *satellite* and not tagged with *Emotion Holder* (*EH*) feature, the word is selected as a potential candidate for topic.

Different unigram and bi-gram context features (word and POS tag level) and their combinations were generated from the training corpus.

[5] English to Indian Languages Machine Translation (EILMT) is a TDIL project undertaken by the consortium of different premier institutes and sponsored by MCIT, Govt. of India.

[5. Evaluation]

The combination of multiple features in comparison with a single feature generally shows a reasonable performance enhancement of any classification system. The impact of different features and their combinations were measured on the development set of 630 sentences and the results are given in Table 1. We added each feature into the active feature list one at a time if the inclusion of the feature in the existing feature set improve the *F-Score* of the system on the development set. The final active feature set has been applied on the test data. During SVM-based training phase, the current token word with three previous and three next words and their corresponding POS along with negation or intensifier were selected as context feature for that word. It is observed that the application of the pruning as a post processing strategy improves the performance of the system significantly.

Information Gain Based Pruning (IGBP): The importance of incorporating the features is examined through Information Gain (*InfoGain*). This decision technique is used to measure the importance of a feature (X) with respect to the class attribute (Y). Formally, information gain of a feature X with respect to a class attribute Y is the reduction in uncertainty about the value of Y when we know the value of X.

$$\textit{InfoGain}(Y;X) = \textit{entropy}(Y) - \textit{entropy}(Y|X)$$

where X and Y are discrete variables taking values $\{x_1, x_2, ..., x_m\}$ and $\{y_1, y_2, ..., y_n\}$ respectively. The Entropy (Y) is defined as

$$\text{Entropy}(Y) = - \sum_{i=1}^{n} P(Y =_{y_i}) \; log_2 \; P(Y =_{y_i})$$

The conditional entropy of Y given X is defined as

$$\text{Entropy}(Y|X) = - \sum_{i=1}^{m} P(X =_{x_j})\text{Entropy}(Y|X =_{x_j})$$

Features with high Information Gain reduce the uncertainty about the class to the maximum. In our experiment on the development set, all the features except causal verbs and transitive dependency relations achieve an Information Gain above 50%. Hence, the information Gain of (50%) has been fixed as Information Gain (*InfoGain*) threshold. The word features (e.g. non-emotional words such as *gather*, *seem*) are filtered from the corpus based on the Information Gain (*InfoGain*) threshold.

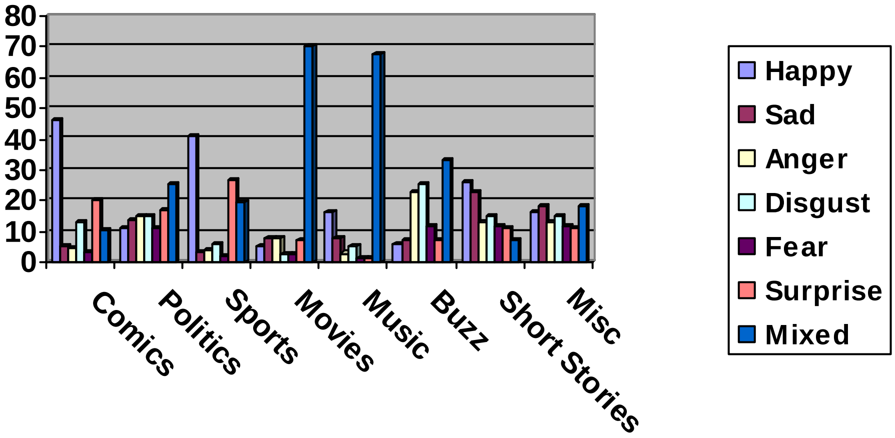
Finally, we use Passonneau's (2004) [24] generalization of Krippendorff's (1980) [25], a standard metric employed for inter-annotator reliability studies. Krippendorff's $\boldsymbol{\alpha}$ is based on a probabilistic interpretation of the agreement of coders as compared to agreement by chance. While Passonneau's innovation makes it possible to apply Krippendorff's $\boldsymbol{\alpha}$ to coreference classification, the probabilistic interpretation of the statistic is unfortunately lost [15].

Generally, the $\boldsymbol{\alpha}$ score aims to probabilistically capture the agreement of annotation data and separate it from chance agreement. The score that we observed for the overall agreement in the baseline system was an $\boldsymbol{\alpha}$ of .53, which is below the generally accepted level, while $\boldsymbol{\alpha}$ for the supervised system is of .67. The $\boldsymbol{\alpha}$ score loses its probabilistic interpretation due to the way that it is adapted to the problem of coreference classification. It is observed that the score of $\boldsymbol{\alpha}$ has been increased rapidly while considering the syntactic, rhetoric and overlapping features. The overlapping features also causes problem in the performance as Bengali is a free phrase order language. But, the application of Named Entities (NEs) reduces the problem of distinguishing user-topic coreference. The overlapping context of emotional expression and topic generates errors but the application of rhetoric knowledge of *nucleus* and *satellite* that is coreferred by locus contributes to separate the overlapping portions. The coreference scores for some important features, their combinations and post-processing steps are shown in Table 1.

[Table 1. Krippendorff's α for different features and post-processing]

Features	Krippendorff's α
Baseline System	0.5344
Supervised System (Lexical Features)	0.3561
Supervised System (Syntactic Features)	0.4002
Supervised System (Semantic Features)	0.3215
Supervised System (Rhetoric Features)	0.4176
Supervised System (Overlapping Features)	0.2345
Supervised System (Lexical+Syntactic)	0.4890
Supervised System (Syntactic+Rhetoric)	0.5012
Supervised System (Syntactic+Semantic+Rhetoric)	0.5201
Supervised System (Lexical+Syntactic+Rhetoric)	0.5421
Supervised System (Lexical+Syntactic+Semantic+Rhetoric+Overlapping)	0.6221
Supervised System + Post-processing	0.6732

The supervised coreference system identifies the users' different emotion on different topics. A single topic is corefered by several users as well as multiple topics are coreferred by single user. This hypothesis aims to generate many to many correspondences among the blog users and topics. The Ekman's six different emotions are plotted for 8 different topics referred by each of the 22 bloggers. The topic based emotion of the bloggers as shown in Figure 2 signifies that the user-topic coreference system performs to generate the emotional views of the bloggers and its dependence on the associated topics.



[Figure 2. Topic based Emotions of the blog users]

[6. Conclusions]

The automatic extraction of emotional expressions, sentential emotion holders and topics from Bengali blog data is done in the present task. From the overall analysis, it is observed that the identification of coreference is helpful in identifying user-topic relations. The handling of metaphors and their impact in detecting sentence level emotion is not considered. Future analysis concerning the time based emotional change can be used for topic model representation.

[References]

[1] F. Choi, Advances in domain independent linear text segmentation, *Proceedings of NAACL,* 2000.

[2] D. Das and S. Bandyopadhyay, Word to Sentence Level Emotion Tagging for Bengali Blogs, *ACL-IJCNLP 2009,* pp. 149-152, Singapore, 2009.

[3] D. Das and S. Bandyopadhyay, Sentence Level Emotion Tagging on Blog and News Corpora, *Journal of Intelligent System (JIS),* vol.19 (2), pp. 125-134, 2010.

[4] D. Das and S. Bandyopadhyay, Emotion Tagging – A Comparative Study on Bengali and English Blogs, *ICON-09,* pp. 177-184, India, 2009.

[5] D. Das and S. Bandyopadhyay, Emotion Holder for Emotional Verbs – The role of Subject and Syntax, *CICLing- 2010*, A. Gelbukh (Ed.), LNCS 6008, pp. 385-393, Romania, 2010.

[6] D.Das and S. Bandyopadhyay, Developing Bengali WordNet Affect for Analyzing Emotion, *ICCPOL-2010,* California, USA, 2010.

[7] D. Das and S. Bandyopadhyay. Labeling Emotion in Bengali Blog Corpus – A Fine Grained Tagging at Sentence Level, *Workshop on 8th Asian Language Resources, COLING-2010*, Beijing, China, 2010.

[8] Ekman Paul, Facial expression and emotion, *American Psychologist*, vol. 48(4), pp.384–392, 1993.

[9] A. Esuli and F. Sebastiani, SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining, *LREC,* 2006.

[10] T. Joachims, Making large-scale support vector machine learning practical, In B. Sch¨olkopf, C. Burges, A. Smola, editor, Advances in Kernel Methods: Support Vector Machines. MIT Press, Cambridge, MA, 1998.

[11] S. Kim and E. Hovy, Extracting opinions, opinion holders, and topics expressed in online news media text. *Workshop on Sentiment and Subjectivity in ACL/Coling*, 2006.

[12] K. Kipper-Schuler, VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia, 2005.

[13] N. Kobayashi, K. Inui, Y. Matsumoto, K. Tateishi, and T. Fukushima, Collecting evaluative expressions for opinion extraction. *IJCNLP*, 2004.

[14] W. C. Mann and S. A. Thompson, Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, *TEXT* 8, 243–281, 1988.

[15] V. Stoyanov and C. Cardie, Annotating topics of opinions, *LREC,* 2008.

[16] V. Stoyanov and C. Cardie, Topic Identification for Fine-Grained Opinion Analysis, *Coling 2008*, pp. 817–824, 2008.

[17] J. Wiebe, T. Wilson, and C. Cardie, Annotating expressions of opinions and emotions in language, *Language Resources and Evaluation*, vol. 1(2), 2005.

[18] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. *In Proceedings of ICDM,* 2003.

[19] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, Narrowing the Social Gap among People Involved in Global Dialog: Automatic Emotion Detection in Blog Posts, *ICWSM, 2007.*

[20] C. Strapparava and R. Mihalcea, SemEval-2007 Task 14: Affective Text, *ACL,* 2007.

[21] Yu Zhang, Z. Li, F. Ren and S. Kuroiwa, A preliminary research of Chinese emotion classification model. *IJCSNS*, **8**(11),127-132, 2008.

[22] Y. Seki, Opinion Holder Extraction from Author and Authority Viewpoints, *In SIGIR 2007. ACM*, New York, 978-1-59593-597-7/07/0007, 2007.

[23] C. Yang, K. H.-Y. Lin, and H.-H. Chen, Building emotion lexicon from weblog corpora, *ACL*, pp. 133-136, 2007.

[24] R. Passonneau, Computing reliability for coreference annotation. *LREC*, 2004.

[25] K. Krippendorff, Content Analysis: An Introduction to Its Methodology. Sage Publications, Beverly Hills, CA, 1980.

[26] V. Ng, and C. Cardie, Improving machine learning approaches to coreference resolution. In Proceedings of ACL, 2002.

[27] W. Soon, H. Ng, and D. Lim, A machine learning approach to coreference resolution of noun phrases. Computational Linguistics, 27(4), 2001.

[28] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys, vol. 34(1), 2002.

[29] C. O. Alm, D. Roth, R. Sproat, Emotions from text: machine learning for text-based emotion prediction. HLT- EMNLP, pp. 579 - 586, Canada, 2005.

[30] S. Bethard, H. Yu, A. Thornton, V. Hatzivassiloglou, D. Jurafsky, Automatic Extraction of Opinion Propositions and their Holders, *In AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications,* 2004.

[31] R.S. Swier and S. Stevenson, Unsupervised Semantic Role Labelling. *EMNLP,* 2004.

[32] S. Banerjee, D. Das and S. Bandyopadhyay, Bengali Verb Subcategorization Frame Acquisition - A Baseline Model. *ACL-IJCNLP-2009, ALR-7 Workshop,* pp. 76-83, Suntec, Singapore, 2009.

[33] S. Banerjee, D. Das and S. Bandyopadhyay, Classification of Verbs – Towards Developing a Bengali Verb Subcategorization Lexicon. *GWC-2010*, pp. 76-83, India, 2010.

[34] D.K. Evans, A low-resources approach to Opinion Analysis: Machine Learning and Simple Approaches, *NTCIR* , 2007.

[35] J. Hu, C. Guan, M. Wang, F. Lin, Model of Emotional Holder, *In Shi, Z.-Z., Sadananda, R. (eds.) PRIMA 2006. LNCS (LNAI)*, vol. 4088, pp. 534–539. 2006.

[36] Y. Kim, Y. Jung, S.-H. Myaeng, Identifying Opinion Holders in Opinion Text from Online Newspapers. *In 2007 IEEE International Conference on Granular Computing*, pp. 699–702, doi:10.1109/GrC.2007.45, 2007.

[37] A. Popescu, and O. Etzioni, Extracting product features and opinions from reviews. *In Proceedings of HLT/EMNLP*, 2005.

[38] A. Ekbal and S. Bandyopadhyay, Named Entity Recognition using Appropriate Unlabeled Data, Post-processing and Voting, *In Informatica Journal of Computing and Informatics*, ACTA Press, 2008.